The Effect of Fundamental Frequency on Mandarin Speech Recognition

Sharlene Liu, Ph.D.[1], Sean Doyle[2], Allen Morris[3], Farzad Ehsani[4]

[1] *Now at Nuance Communications* *sliu@nuance.com*
[2] *Now at General Magic* *seand@genmagic.com*
[3] *Now at Soft GAM's Software* *gam3@acm.org*
[4] *Now at Sehda* *farzad@sehda.com*

ABSTRACT

We study the effects of modeling tone in Mandarin speech recognition. Including the neutral tone, there are 5 tones in Mandarin and these tones are syllable-level phenomena. A direct acoustic manifestation of tone is the fundamental frequency (f0). We will report on the effect of f0 on the acoustic recognition accuracy of a Mandarin recognizer. In particular, we put f0, its first derivative (f0′), and its second derivative (f0″) in separate streams of the feature vector. Stream weights are adjusted to investigate the individual effects of f0, f0′, and f0″ to recognition accuracy. Our results show that incorporating the f0 feature negatively impacted accuracy, whereas f0' increased accuracy and f0'' seemed to have no effect.

1. INTRODUCTION

In contrast to most European languages, Mandarin Chinese uses tones for lexical distinction. A tone occurs over the duration of a syllable. There are 4 lexical tones and 1 neutral tone. Homonym confusability in Mandarin is very high, and the situation is compounded by not distinguishing among the 5 tones.

An effective Mandarin speech recognizer thus needs to be able to recognize the 5 tones in addition to the usual phonetic inventory of the language. The most direct acoustic manifestation of tone is fundamental frequency (f0). In this paper, we report our experiments with incorporating f0 and its first and second derivatives (f0′, f0″) into the feature vector. We will present the collective effect of all 3 f0 features together as well as the individual effects of each of the features. We put f0, f0′, and f0″ into 3 separate streams with the freedom to individually turn on or off the effect of a stream. Thus, we are able to explore the individual contribution of each stream toward recognition accuracy.

Others have worked on incorporating tone into Mandarin speech recognition. Liu et al. [1] used f0 features to do a large vocabulary continuous Mandarin dictation application. His results will be reported in Section 4. Lyu et al. [2]'s approach differs in that they built a system in which speech is pre-processed to hypothesize syllable boundaries. These hypothesized syllables are then used in a parallel network to identify the base syllable and tone.

In this paper, we explore the effect of f0 features toward accuracy. We describe the database in Section 2. In Section 3, we present a baseline experiment involving no f0 features. In Section 4, we describe a fast and accurate f0 tracker used in our experiment involving f0 features. In Section 5, we look at the individual contributions of each f0 feature toward accuracy by varying stream weights. Section 6 concludes with some ideas on how one might better model Mandarin tones.

2. DATABASE

The speech data were recorded in China using a head-mounted, uni-directional, noise-canceling microphone and sampled at 16 kHz. For the experiments reported in this paper, 54 speakers' data representing about 20 hours of speech were used. Five-sixths of the speech were recorded in a quiet office environment; however, one-sixth of the data were recorded in a noisy office environment. The utterances were continuous, read speech.

2.1. Sentence Corpus

Thousands of sentences were selected and edited for the recording prompts. These sentences came from newspapers, magazines, and novels published in China. The diversity in sentences guarantees a wide range of phonetic, lexical, and prosodic content. Sentences were chosen based on readability and phonetic coverage.

2.2. Speaker Demography

Approximately half of the speakers were female and half male. By age, 65% were 18-30 years old, 15% were 31-40 years old, 14% were 41-50 years old, and 6% were 50-65 years old.

China is the most populous country in the world. Many languages and accents exist. Thus, it is of utmost importance to choose speakers from diverse geographic and lingual origins. The Beijing accent was focused upon, however, because the official accent is found in Beijing. 30% of the speakers were born in Beijing. 20% of the speakers identified Mandarin as their native language; 80% used Mandarin at home. Other speakers came from other parts of China, including Shanghai, Guangdong, Fujian, and Sichuan. Of course, all the speakers spoke Mandarin fluently

2.3. Utterance Verification

Each utterance was verified for accuracy. Differences in accents presented many ambiguities when verifying. The rule used was that if the variation in speaking was due to a dialectal accent, then the utterance was considered "good"; however, if

the variation was unintended, then the utterance was considered “bad”. For example, many southern accents do not have retroflexion, so the phonemes /ch, sh, zh/ were pronouced as [c, s, z], respectively. Sentences pronouced by southerners without retroflexion were "good" but the same sentences pronounced by northerners without retroflexion were "bad". Another example is the retroflexed ending: northerners pronouce many words with a retroflexed "er" at the end. Sentences pronounced in this way by northerners were considered “good”.

3. BASELINE EXPERIMENT

3.1. Phonetic Inventory

There are 32 phonemes in Mandarin, grouped below by broad phonetic class:

- Vowels a, E, i, ih, o, u, U, u:
- Semivowels r, l, w, y
- Nasals m, n, ng
- Fricatives f, s, sh, x, h
- Affricates z, c, zh, ch, j, q
- Stops b, d, g, p, t, k

The acoustic model inventory consisted of tonal vowels and atonal consonants. There were thus 8 vowels x 5 tones/vowel = 40 tonal vowels. Adding 24 consonants made a total of 64 monophone models to train. The consonants were specified to be atonal because most of the tone information is carried in the vowel portion of a syllable.

3.2. Training

Monophone models were built using Entropic’s HMM ToolKit (HTK) [3]. In the baseline experiment, we used 13 Mel-Frequency Cepstral Coefficients (MFCCs) and their first and second derivatives, making a feature vector that was 39 elements long. f0 features were not used for the baseline. Thus, the tonal vowels were distinguished by the MFCCs, which are not expected to be helpful in distinguishing tone since f0 information is thrown away in the computation of the MFCCs.

Models were built using a flat-start approach in which each model was seeded by the global means and variances calculated from all the training data. The training data consisted of 48 speakers’ speech and approximately 14,000 utterances. The silence and short pause models were refined with iterations of the Baum-Welch re-estimation algorithm. Initially, single-gaussian monophone models were built. Then from those, gaussians were added to achieve 16-mixture models.

3.3. Testing

A word in Mandarin is not clearly defined as it is in western languages. Words are not written with spaces in between them. Word boundaries are ambiguous and often there are more than one way to parse a sentence into words. Every syllable in Mandarin is also a word; thus a multi-syllablic word can be broken up into multiple single-syllable words. For example, a word like “pian1jian4” can be broken up into 2 words, “pian1 jian4”, where the digit indicates tone. If the recognizer recognizes the word as the 2 words “pian1 jian4” instead of the 1 word “pian1jian4”, the acoustic scoring should still consider it as correct. Toward that end, word-level recognition was performed but the results were scored on a syllable basis.

In order to nullify the effects of the language modeling and concentrate on acoustic modeling, an all-word parallel language network was constructed such that the language model was limited to equal unigram probabilities for all words.

A small test set of 6 speakers and a total of about 2000 utterances was used. The test sentences were hand-parsed into words. The word-level recognition task using syllable-level scoring gave a baseline accuracy of 61.61%, as shown in Table 1. This accuracy should not be taken as the best accuracy achievable for the model topology and training method used. Indeed, we increased the training data by a factor of 5 and saw the absolute accuracy jump up to 70%. No doubt, adding even more training data and training triphone instead of monophone models would cause a similar jump in accuracy. The accuracy is, however, important as a baseline against which to compare the subsequent experiment which uses f0 features.

Experiment	Accuracy	Error reduction
Baseline	61.61 %	--
Baseline + f0 features	64.04 %	6.3 %

Table 1: Results of baseline experiment and f0 experiment. Adding f0 features reduces error rate by 6.3%.

4. TONAL MODELS WITH f0

Next, we turn to the effect of adding f0 parameters to the feature vector. As stated in Section 1, f0 is a direct acoustic manifestation of tone. Thus, adding f0 and its derivatives to the feature vector should improve the accuracy on tonal tasks. Strictly speaking, tone is a syllable-level phenomenon and simply adding f0 features to the phonetic models is not an accurate way of recognizing tone. A syllable can have more than one vowel in its nucleus, and the f0 track is different depending on whether the vowel is at the beginning or the end of the syllable. However, as a simplification, this method does reveal the usefulness of f0 and its derivatives.

4.1. f0 Tracking Algorithm

Talkin’s f0-tracking algorithm [4] was used to derive f0 values for each frame. This algorithm uses a two-pass strategy. In the first pass, it calculates, using rough time parameter values, the short-time cross-correlation between a windowed segment of the speech waveform and a time-shifted version of itself. The normalized peak of the cross-correlation is taken to be a likely vicinity where a pitch pulse has occurred. In the second pass, a

finer time window is used in the cross-correlation calculation to more accurately define the time of the pitch pulse in the vicinity of a peak found in the first pass. Once the pitch pulse candidates are identified, a dynamic programming stage across syllabic time scales is used to smooth out discontinuities in the f0 track caused by pitch doubling or halving.

The f0-tracking algorithm can be computed in real-time. The first and second passes can be computed in parallel. The dynamic programming stage in the post-processing can be limited to 100 ms, which is short enough to seem real-time to a user.

Where speech is unvoiced, f0 is undefined; we set it to zero to give it a fixed value. In order for the derivatives to not be discontinuous at voiced/unvoiced boundaries, we forced the first derivative to be zero for the first 2 frames into a voiced region on the left and right boundaries. Similarly, we forced the second derivative to be zero for the first 2 frames into the non-zero first derivative region.

4.2. Feature Vector Extension

f0, f0′, and f0″ were appended to the 39-element MFCC feature vector as 3 independent streams. The resulting feature vector thus comprised 4 independent streams: MFCCs + derivatives, f0, f0′, and f0″. The streams weights were set to [1 1 1 1], i.e. equal weighting for each stream. This notation for stream weights refers to the weights for MFCCs + derivatives, f0, f0′, and f0″, respectively. The model topology forces each stream to be statistically independent of the others. Such an assumption is approximately correct for the MFCCs vs. f0 insofar as the glottal source is independent of vocal track shaping.

4.3. Training

Just as in the baseline experiment, 64 monophone models, consisting of 40 tonal vowels and 24 consonants, were trained. The models were again mixed up to 16 gaussians using the same training set as in the baseline experiment.

The stream topology was incorporated into the training phase of model building. The contribution of each stream's output probability was multiplied together. Gaussian mixtures within each stream were added together. The output probability distribution is described by

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} N(o_{st}; \mu_{jsm}; \Sigma_{jsm}) \right]^{\gamma_s} \qquad (1)$$

where

j = state index
t = time index
s = stream index
S = number of streams
m = gaussian mixture index
M = number of mixtures
b = output distribution
o = observation vector
c = stream weight
$N(o; \mu; \Sigma)$ = gaussian distribution with mean μ and covariance Σ
γ = stream weight

4.4. Testing

For the recognition test, the test set used in the baseline experiment was used here. The same parallel word language network was applied, and word-level recognition with syllable-level scoring was again performed. By adding the f0 information to the feature vector, we realized a 2.43% increase in absolute accuracy, which represents a 6.3% relative error reduction. This improvement is a rather significant result, indicating that the use of f0 is unquestionably beneficial. A similar experiment by Liu et al. [1] on the Mandarin Call Home corpus also indicated that f0 features do help recognition, although their reduction in relative error using f0 features was only 1%.

5. STREAM WEIGHTS

We now explore the effect of stream weights. In Section 4, the stream weights used were [1 1 1 1], i.e. the MFCCs+derivatives, f0, f0′, and f0″ were weighted equally. The question arises: how much do f0, f0′, and f0″ each contribute to the improvement in accuracy? By varying stream weights, we can explore the effect of each individual stream. In this section, we explore the effect of f0 and f0″ by setting some of the weights to 0. According to Eq. (1), a stream weight of 0 nullifies the effect of the stream. The stream can then be dropped from the feature vector, saving both CPU time and memory.

5.1. Baseline: Single Gaussian Models

In order to shorten training time, we constructed single-gaussian monophone models rather than the 16-gaussian mixture models described in Sections 3 and 4. Thus, a new baseline experiment was conducted. The training procedure is the same as before, using the same amount of training data, but this time we stopped at single gaussians and did not add mixtures. Again, word-level recognition and syllable-level scoring were performed. The resulting accuracy was 35.28%, as shown in Table 2.

Experiment	Accuracy	Error reduction
Baseline [1 1 1 1]	35.28%	--
Nullify f0 [1 0 1 1]	37.53%	3.47%
Nullify f0″ [1 0 1 0]	37.53%	3.47%

Table 2: Results of stream weight experiments. Stream weights are shown in []'s. The first row is the baseline. The second row shows the effect of nullifying f0. The third row shows the effect of nullifying f0″.

5.2. Effect of f0

The normal f0 range of individual adult speakers easily spans the range 100-300 Hz. Even in normal speech, f0 for women can exceed 300 Hz, while for men it can fall below 100 Hz. Unless f0 is normalized, it very likely adds little toward accuracy improvement. Without adaptation or *a priori* information on the speaker's f0 range, normalizing f0 is impractical. In the baseline experiment of Section 5.1, we did not normalize f0. f0, then, is not expected to significantly affect accuracy. Thus, an experiment using stream weights [1 0 1 1] was conducted to explore this hypothesis.

Single gaussian monophone models were trained as before, but this time with stream weights [1 0 1 1]. Recognition accuracy for these weights turned out to be 37.53 %, representing 3.47% error reduction, as shown in Table 2. As expected, the accuracy was improved by nullifying the f0 stream.

5.3. Effect of f0″

In addition to exploring the effect of f0, one may question the utility of f0″. Canonical tones in Mandarin are characterized by piece-wise linear segments of f0 as a function of time. Because they are piece-wise linear, the second derivative of f0 should be 0 except where there are discontinuities in f0. Canonically, a discontinuity within a syllable can only occur for tone 3, for which f0 as a function of time is V-shaped. However, because we are building phonetic models without regard to phone position in a syllable, values of f0″ are not likely to be consistent across phones. Of course, syllable-boundary effects also have discontinuities in f0, but again these effects are not likely to show regularity at the phoneme level. The stream weighting [1 0 1 0] is therefore of interest.

After models were trained and tested, we realized a recognition accuracy of 37.53%, as shown in Table 2. This signifies that f0″ with the presence of f0′ neither improves nor hurts the accuracy.

6. CONCLUSION

The fundamental frequency was used to model tones in Mandarin Chinese. f0 and its derivatives were shown to significantly improve the recognition accuracy. With streams, we were able to isolate the effects of f0, f0′, and f0″. We showed that f0 is not a useful feature when it is not normalized. This is because of the wide range of f0 in speakers. We also showed that f0″ does not contribute much to recognition accuracy when doing phoneme-level modeling. The reason is that canonical tones are piece-wise linear over a syllable, so f0″ is canonically 0 except at break points in f0.

In future work, more effective tone modeling should be pursued. Tones are syllable-level phenomena. In this paper, we modeled tones at the phonemic level because this technique is simple to incorporate into the existing phoneme-based acoustic models. Eventually, a syllable-level model should be pursued. Such modeling can be achieved through syllable-based acoustic models. Since the number of tonal syllables in Mandarin is limited to 1345 syllables, the amount of training data to build syllable models is not prohibitive. Further efficiencies can be realized by splitting the syllable into its initial and final components.

Syllable models also allow context-dependent tone models to be built easily. Thus we can imagine building *tri-syllable* models, with left and right context dependency of not only neighboring phonemic effects, but neighboring tone effects as well. Tones influence each other dramatically, as shown in a study by Xu [5], so context-dependent tone models are indispensible in Mandarin speech recognition.

7. ACKNOWLEDGEMENTS

We would like to thank Entropic Inc. for providing the software and data used in this research. Special thanks to Dr. Steve Young for his insights.

8. REFERENCES

1. F. Liu, M. Picheny, P. Srinivasa, M. Monkowski, J. Chen, "Speech Recognition on Mandarin Call Home: A Large Vocabulary, Conversational, and Telephone Speech Corpus," *Proc. ICASSP*, 1996, Vol. 1, pp. 157-160.

2. R. Lyu, L. Chien, S. Hwang, H. Hsieh, R. Yang, B. Bai, J. Weng, Y. Yang, S. Lin, K. Chen, C. Tseng, L. Lee, "Golden Mandarin (III) – A User-Adaptive Prosodic-Segment-Based Mandarin Dictation machine for Chinese Language with Very Large Vocabulary," *Proc. ICASSP*, 1995, pp. 57-60.

3. S. Young, J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory, Cambridge, England, 1997.

4. D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*, W.B. Kleijin, K.K. Paliwal, Editors, Elsevier Science B.V., 1995, pp. 91-114.

5. Y. Xu, "Contextual Tonal Variations in Mandarin," *J. Phonetics*, 1997, Vol. 25, pp. 61-83.